Second rewiewer:

The main ones are:

i) the explanation of the drivers of upper plate extension is hard to follow and I think, in parts, incorrect;

ii) there's a lack of justification/testing for the rheological properties of the upper plate (which are obviously important for whether a plate breaks or remains intact);

iii) the model velocities are extremely high but this is not discussed (but would affect upper plate extension significantly).

Main points:

Extension mechanism: You propose that extension can be triggered by either an extensional force in the plates, or subduction-induced flow beneath the plates, and rule out the first option (i.e., an extensional force, which you propose is related to the speed difference of the plates). This is misleading because extension will always be due to extensional normal stresses within the (pre-extended) plate. A more accurate way of framing this is whether this extensional normal stress is due to a horizontal force transmitted from the plate boundary (e.g., due to rollback) or, as you say, tractions from mantle flow. I don't disagree that basal tractions are the dominant control, just with how you are describing the different stress components/framing the physical problem. Also, about this mantle flow contribution, you state that it's dominantly the vertical (not horizontal) flow. But, prior to spreading, it's horizontal flow that produces basal tractions on the base of the near-flat lithosphere (which, yes, does ultimately originate from vertical flow that has been deflected horizontally). Yes, we agree that an extensional is always due to extensional normal stresses within the plate. Maybe we didn't describe it clearly but what we want to emphasise is that the extension is triggered by the decreasing strength of the OP but not increasing extensional force. We have edited the text in Discussion and hope it makes our point clearer.

There are a lot of modeling studies that delve into what dictates the upper plate stress state in dynamic subduction models (Capitanio et al., 2010, Tectonophys.; Schellart and Moresi, 2013, JGR; Holt et al., 2015, GJI; Dal Zilio et al., 2018, Tectonophys.) and carefully consider the relationship between sub-plate flow, basal tractions, and lithospheric stress. You cite some of these in passing; I recommend integrating the perspectives of these previous studies to present a clearer view of the forces in your upper plate and how they vary between models. An improved force description (Section 4.1), and an incorporation of this into Section 4.2, should make the discussion much clearer. We have edited our discussion. Also, in Figure 10, you plot integrated stress profiles and show that extension/spreading is triggered in a compressional region; this cannot be correct so should be sorted out (either by outputting more timesteps or plotting a zoom-in of the stress field for the corresponding timesteps to check your stress integration is correct). Sorry, Figure 10 in the original manuscript was in error, we wrongly marked the location of Extension in EH mode. Currently, the Extension is triggered in an extensional region, which is consistent with intuition.

Upper plate properties: You are investigating a balance between the forces driving the extension and the plate strength (i.e., extension when driving forces > strength) and so the

imposed strength of the upper plate is very important. I therefore recommend more discussion (or additional tests) of the parameters you choose that dictate this. It looks as though the non-extended (or pre-extended) lithospheric strength is dictated by the maximum imposed viscosity (10^{25} Pa s) and that extension occurs once the stress > the plastic yield stress (2 MPa + 0.2 * Pressure). And while the hot zone will also lower the viscosity and so reduce the strength, you do not specify by how much. Given the importance, I recommend more discussion about your plastic yielding parameterization. What does this yielding viscosity represent? What dictated/justifies the parameter choices? You might run some tests to show that your first order findings do not depend on some of these choices too much. The plastic yielding parameterisation is very widely used, and can be thought to represent frictional failure. The choice of the 'friction coefficient', as mentioned in the text, falls between the estimates of Byerlee's law, and the much lower estimates of the friction coefficient at the SP/OP interface by other researchers. Since plate weakening dominates over increasing stresses, and plate weakening starts at the base of the OP, the critical rheologies are the viscous ones dominant at the base of the OP. Therefore, while the yielding will play a role in the total strength, and therefore how much stress is required to rift the plate, this will be secondary to the role of the other rheologies. We feel that if rheology parameters were to be investigated, they should all be investigated. That is beyond this work, but we agree is very worthy of investigation, and we have undertaken a separate project to this effect. We have also run some models increasing the maximum composite rheology and found that it makes no difference. The decrease in viscosity from the hot zone varies spatially according to the specific increase in temperature at that location, as given by the composite rheology. Therefore, it is not straightforward to specify by how much the viscosity is decreased - it is different within any case, and more so between cases. This is best done through figures and we have added the viscosity drop due to the hot zone in figure 4. In the caption of Figure 4: 'Figures (b) (d) (f) (h) (j) show the viscosity field and plots of the initial vertical viscosity at the hot region centre in the 0 Ma figure of each model represented by the red lines, while the black lines represent the viscosity without the hot region. The two lines in different colours show the change in viscosity resulting from emplacing a hot zone.'

Model velocities: You don't quote model convergence rates or trench motion rates but describe the slab hitting 660-km in 4 Myrs. This corresponds to very high sinking velocities (> 10 cm/yr) and very high mantle flow velocities (Figure 11). These velocities are likely very important in setting the stress in your models, and hence when extension/spreading occurs. I think should be discussed or, at the very least, explicitly pointed out. Yes, high sinking velocities (and therefore convergence velocities) do exist in these models for a short time as subduction starts, and drop significantly once the slab enters the lower mantle. Thank you for encouraging us to add some discussion on this point in the manuscript, added at lines 133-136. 'We note that the SP sinking velocity in this short time period increases to a very high value (with local peak of > 10 cm/yr, and a lower peak when averaged over 1 Myr [geologically observable timeframe]), as expected, as the length of subducting slab increases, before decreasing to nearly steady values of < 2 cm / yr once the slab approaches and enters the more viscous lower mantle.' This significant variation in sinking velocity is seen in many researchers work when they consider this early stage of subduction (Lei and Davies, 2023; Garel et al., 2014; Hall et al., 2003).

Line-by-line:

- L9: Is it a competition of "thermal weakening" between these two regions? Or is where has the largest extensional stress relative to the strength (which, at a certain location, is reduced due to the arc)? Yes, of course where extension occurs is strictly given by the statement included in the reviewer's second question. The extra weakening process though is the thermal weakening at both the hot-region and the far-field location. We have therefore decided to retain the statement since this is the main way that we describe the process in the article. We thank the reviewer for forcing us to consider this point.
- 22: Lots of studies looked at the controls on upper plate stress, so they did (albeit indirectly): e.g., those mentioned above. We have edited the text to now mention "many … previous models …". (Lines 20-30 in the revised manuscript.)
- 27-31: This passage summarizes the motivation/novelty very nicely (but I'm not sure what the Bettina reference is attached to). Dropped this citation off.
- ~47-49: Is it where the properties are changing the most quickly (as you write)? Or just spatial gradients at a given timestep? Yes, this is correct. Also, which properties do you use to refine? We refine at high spatial gradients of viscosity, second invariant of strain-rate, temperature and weak zone phase amount. We add the following sentence to address these points 'The grid adapts throughout the simulation, keeping the finest resolution where the spatial gradients of fields (viscosity, second invariant strain-rate, temperature and weak zone phase amount) are highest.' (Lines 54-55 in the revised manuscript.)
- 57: "around 194" -> "194". Edited.
- 66: What is this "prescribed depth"? The same as the depth of the plate bottom, which varies in different OP ages. We have edited in the text: 'The temperature at a chosen distance (from 100 to 1050 km) from the trench is increased by a certain degree (the degree varying from 25 to 800 degrees) vertically from the mantle depth (which varies in various OP ages) to the surface.' (Lines 70-71 in the revised manuscript.)
- Equation 4: You call p both lithostatic and dynamic pressure. I think it's the "full" pressure (i.e., the sum of these two). We use the capital 'P' to represent the lithostatic pressure in Equation 9 and the small letter 'p' to represent the dynamic pressure in Equation 4.
- ~95: Where does this simplified parameterization of Peierls creep come from? Ref(s) needed. Added the description and reference in the text at Lines 116-117 in the revised manuscript.
- Equation 9: 2nd tau_y should be a tau_0. Edited.
- 113: Viscous dashpots in series (i.e., the strain rates sum) not parallel. See Schmeling et al. (2008, PEPI). Thank you. The sentence has been edited to mention strain rate and be clearer. 'This is assuming that the strain rates of all 4 deformation processes sum, like viscous dashpots in series (Schmeling et al. 2008)'. (Lines 120-121 in the revised manuscript.)
- 160: about one-tenth -> one-tenth. Edited at Line 124 (revised version).
- 129: Difference between eroded and thinning? There is no fundamental difference between 'eroded' and 'thinning' in our work, so we have unified them to 'thinning' in the text (Line 137 and 140 in the revised version).

- Sect. 3.1: I think this description of the modes is quite confusing and can be simplified. Particularly the no extension vs. extension. E.g., on L139, you say that the state before complete thinning is classified as No Extension; but, on L141, you suggest that No Extension also corresponds to some thinning. I would just try and simplify this. Thank you for this comment. We have taken the opportunity to look at this again. We believe that having 2 states of extension is the simplest, leading to the 3 modes. We have re-read our description of our definition of these states at the first paragraph at the start of section 3.1 and feel that it is very clear. We state clearly that thinning is included as No Extension, while once it reaches complete thermal thinning and rifting it is Extension. We prefer to leave this classification but have changed 'Complete Thinning' to 'Complete Thermal Thinning and Incipient Extension' to make less confusion. We thank the reviewer for encouraging us to look at this again.
- Figure 2: I would add a length scale to the figures, particularly as you are talking about trench-extension distances. And I don't understand the stress units. We have added a length scale to all relevant figures and changed the stress unit to 'Pa'.
- 164-165: But is extension at the hot region (HR) always "close to the trench"? Because you are moving the HR quite far away, so it's quite far away? Find this confusing in your definition of EF vs. EH. No, HR can be far away from the trench. Our EF mode means the extension occurs both 'not at the HR' and 'far away from the trench'. There are some models in which the HR is far away from the trench and the back-arc extension occurs at the HR, these cases are called EH as well.
- 187: SP velocity, convergence rate, or slab sinking velocity? We thank the reviewer for pointing out the need for further clarification. This is in fact the vertical slab sinking velocity. The text is updated appropriately in Lines 199-200.
- 198-199: I think a weak OP just provides less horizonal resistance to rollback. E.g., single slabs models without OPs (e.g., old subduction models such as Enns et al. [2005, GJI]) always have high rollback as they are basically weak-OP endmembers. Thank you for the suggestion, we have edited the text in Line 211 (revised version).
- 230: The speed differences within the upper plate (as they produce horizontal normal strain rate) not the speed difference between the plates. We have edited the text in Lines 240-242: 'In the first mechanism the extensional stress is generated from the speed difference between the trench and the OP, while the second arises from variable shear stresses at the base of the OP.'
- 237-239 and Figure 10: It's hard to see what the issue is – Can you also show zoomed-in plots of the horizontal stress field at equivalent times, e.g., as the Schellart & Moresi paper does. Perhaps you are outputting the stress after the extension has occurred? Instead of right before. Sorry, we mislocated one star marker in Fig.10 in the original manuscript. We have edited it in the figure and updated the relevant text (Line 253-259 in revised manuscript).
- 247: As mentioned, I think the upwelling flow would weaken the upper plate via basal drag (i.e., these two things are one mechanism). Unless you are talking about after extension, and during spreading, when the upwelling would sustain spreading. But it's not really clear from the text explanation. Yes, we do not mean to ignore the importance of the basal drag but wanted to emphasise the thermal weakening caused by the upward flow. We have edited the text at Line 267-269 by adding a comment to make clear that the basal drag can also weaken the OP, and emphasise that both components ultimately result from upwelling flow. 'The trench-ward horizontal flow

(X component) produces basal drag by the velocity gradient (Figure 13), which facilitates an Extension by producing the extensional force and weakening the OP, whereas the Y component of the flow also encourages the thermal weakening of the OP. The upwelling flow, the cause of both components of flow, is always in a similar direction.'

- Figure 11: is the velocity scale really up to 105 cm/yr?! Or is this a typo? No, it is not a typo, it is up to this value, but only very briefly. We have discussed these high velocity values earlier in the response and added a comment in the manuscript.
- 265-266: I think incorporating the results of these studies (and Capitanio et al., 2010; Schellart and Moresi, 2013, etc.) would make this discussion a bit clearer from a mechanism point of view. Those studies outline where (and how much) extension we get in OPs; you are then effectively combining this logic (about driving forces) with a weak zone of various magnitudes. Thank you. We have changed the text to take up your suggestion and added your suggested reference. 'There have been many studies investigating back-arc extension and some have emphasised the balance of strength and forces that lead to extension and shown that the location of extension in the OP can be related to the flow cell in the mantle wedge. In our work here we similarly can find extension in the same flow cell controlled location, but emplacement of a hot region can sufficiently weaken the OP to change the location of extension to this weakened region.'
- 271: What is meant by "thermal weakening"? We have added a description in the text 'reduced viscosity due to increased temperature'. (Line 300 in revised manuscript)
- 274: In 2-D, the mantle wedge flow will be more strongly controlled by the convergence or slab sinking rate. See the classic corner flow papers by McKenzie, Tovish and Schubert, etc. Yes, agreed. We have edited the text to reinforce the importance of convergence rate (Line 303 in revised manuscript).
- 286: Upwelling "intrusion" is maybe confusing when talking about the drivers of extension – the intrusion only happens after extension has occurred and spreading has created the space. Edited 'intrusion' to 'flow'.
- 323-329: What are the references for these back-arc basin ages? Sdrolias and Muller? Also see Clarke, Stegman, Muller (2008, PEPI). We have added some references to these back-arc basin ages (Lines 354-363 in revised manuscript).
- Section 4.4: In this limitation sections, I recommend: i) avoiding the double list (i.e., two layers of numbering); ii) including refences to studies that have considered these complexities; iii) organizing it more intuitively (e.g., going from simplifications that you think are the most important to those that are the least). We have preferred to keep the double list (but changing one set of numbers to letters) to group the limitations. Within this constraint we have tried to go from the most important to the least important. We have added a few references that have considered these complexities (Lines 381-384 in revised manuscript).
- 369: A higher slab sinking/convergence rate (which likely coincides with a high trench retreat rate). Thank you. We have clarified this by adding 'through a higher slab sinking rate' to the text (Line 406 in revised manuscript).
- 370: What does you will share your models "upon reasonable request" mean? Consider sharing your input files in an open online repository. If anybody contacts me, the

corresponding author, then I will share with them any model outputs they are interested in.